Sampling, feasibility, and priors in Bayesian estimation

Alexandre J. Chorin[1,2], Fei Lu[1,2], Robert N. Miller[3], Matthias Morzfeld[1,2], Xuemin Tu[4]

[1]Department of Mathematics, University of California, Berkeley, CA
[2]Lawrence Berkeley National Laboratory, Berkeley, CA
[3]College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR
[4]Department of Mathematics, University of Kansas, Lawrence, KS

DEDICATED TO PETER LAX, ON THE OCCASION OF HIS 90th BIRTHDAY, WITH FRIENDSHIP AND ADMIRATION

Abstract

Importance sampling algorithms are discussed in detail, with an emphasis on implicit sampling, and applied to data assimilation via particle filters. Implicit sampling makes it possible to use the data to find high-probability samples at relatively low cost, making the assimilation more efficient. A new analysis of the feasibility of data assimilation is presented, showing in detail why feasibility depends on the Frobenius norm of the covariance matrix of the noise and not on the number of variables. A discussion of the convergence of particular particle filters follows. A major open problem in numerical data assimilation is the determination of appropriate priors; a progress report on recent work on this problem is given. The analysis highlights the need for a careful attention both to the data and to the physics in data assimilation problems.

1 Introduction

Bayesian methods for estimating parameters and states in complex systems are widely used in science and engineering; they combine a prior distribution of the quantities of interest, often generated by computation, with data from observations, to produce a posterior distribution from which reliable inferences can be made. Some recent applications of these methods, for example in geophysics, involve more unknowns, larger data sets, and more nonlinear systems than earlier applications, and present new challenges in their use (see, e.g. [7,15,19,43,45] for recent reviews of Bayesian methods in engineering and physics).

The emphasis in the present paper is on data assimilation via particle filters, which requires effective sampling. We give a preliminary discussion of importance sampling and explain that implicit sampling [3,12,13,32] is an effective importance sampling method. We then present implicit sampling methods for calculating a posterior distribution of the unknowns of interest, given a prior distribution and a distribution of the observation errors, first in a parameter estimation problem, then in a data assimilation problem where the prior is generated by solving stochastic differential equations with a given noise. Conditions for data assimilation to be feasible in principle and their physical interpretation are discussed (see also [11]). A linear convergence analysis for data assimilation methods shows that Monte Carlo methods converge for many physically meaningful data assimilation problems, provided that the numerical analysis is appropriate and that the size of the noise is small enough in the appropriate norm, even when the number of variables is very large. The keys to success are a careful use of data and a careful attention to correlations.

A very important open problem in the practical application of data assimilation methods is the difficulty in finding suitable error models, or priors. We discuss a family of problems where priors can be found, with an example. Here too a proper use of data and a careful attention to correlations are the keys to success.

2 Importance sampling

Consider the problem of sampling a given probability density function (pdf) f on the computer, i.e., given a probability density function (pdf) for an m-dimensional random vector, construct a sequence of vectors $X_1, X_2, \dots,$ which can pass statistical independence tests, and whose histogram converges to the graph of f, in symbols, find $X_i \sim f$. We discuss this problem in the context of numerical quadrature. Suppose one wishes to evaluate the integral

$$I = \int g(y) f(y) dy,$$

where $g(y)$ is a given function and $f(y)$ is a pdf, i.e., $f(y) \geq 0$ and $\int f\, dy = 1$. The integral I equals $E[g(x)]$, where x is a random variable with pdf f and $E[\cdot]$ denotes an expected value. This integral can be approximated via the law of large numbers as

$$I = E\left[g(x)\right] \approx \frac{1}{M} \sum_{i=1}^{M} g(X_i),$$

where $X_i \sim f$ and M is a large integer (the number of samples we draw). The error is of the order of $\sigma(g(x))/M^{1/2}$, where $\sigma(\cdot)$ denotes the standard deviation of the variable in the parentheses. This assumes that one knows how to find the samples X_i, which in general is difficult. One way to proceed is importance sampling (see, e.g., [8, 24]): introduce an auxiliary pdf f_0, often called an importance function, which never vanishes unless f does, and which one knows how to sample, and rewrite the integral as

$$I = \int g(y) \frac{f(y)}{f_0(y)}\, f_0(y) dy = E\left[g(x^*) w(x^*)\right]$$

where the pdf of x^* is f_0 and $w(x^*) = f(x^*)/f_0(x^*)$. The expected value can then be approximated by

$$I \approx \frac{1}{M} \sum_{i=1}^{M} g(X_i^*) w(X_i^*) \tag{1}$$

where $X_i^* \sim f_0$ and the $w(X_i^*) = f(X_i^*)/f_0(X_i^*)$ are the sampling weights. The error in this approximation is of order $\sigma(gw)/M^{1/2}$ (here the standard deviation is computed with respect to the pdf f_0). The sampling weights make up for the fact that one is sampling the wrong pdf.

Importance sampling can work well if f and f_0 are fairly close to each other. Suppose they are not (see figure 1). Then many of the samples X_i (drawn from the importance function) fall where f is negligible and their sampling weight is small. The computing effort used to generate such samples is wasted since the low-weight samples contribute little to the approximation of the expected value. In fact, one can define an effective number of samples [2, 15, 26, 48] as

$$M_{\text{eff}} = \frac{M}{R}, \quad R = \frac{E\left[w^2\right]}{E\left[w\right]^2},$$

so that in particular, if all the particles have weights $1/M$, $M_{\text{eff}} = M$. The effective number of samples approximates the equivalent number of independent, unweighted samples one has after

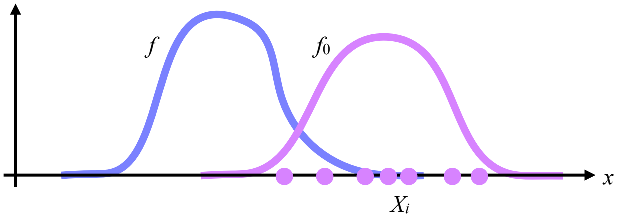

Figure 1: Poor choice of importance function: the importance function (purple) is not close to the pdf we wish to sample (blue).

importance sampling with M samples. Importance sampling is efficient if R is smaller than M. For example, suppose that you find that one weight is Rmuch larger than all the others. Then $R \approx M$, so that one has effectively one sample and this sample does not necessarily have a high probability. In this case, the importance sampling scheme has "collapsed". To find f_0 that prevents a collapse, one has to know the shape of f quite well, in particular know the region where f is large. In some of the examples below, the whole problem is to estimate where a particular pdf is large, and it is not realistic to assume that this is known in advance.

As the number of variables increases, the value of a carefully designed importance function also increases. This can be illustrated by an example. Suppose that $f = \mathcal{N}(0, I_m)$, where I_m is the identity matrix of order m (here and below we denote a Gaussian with mean μ and covariance matrix Q by $\mathcal{N}(\mu, Q)$). To sample this Gaussian we pick a Gaussian importance function $f_0 = \mathcal{N}(0, \sigma^2 I_m)$, $\sigma > 1/2$. The importance function has a shape similar to that of the function we wish to sample and f_0 is large where f is large (for moderate σ). Nonetheless, sampling becomes increasingly costly as the number of components of x increases. We find that

$$w(x) = \sigma^m \exp\left(-\frac{1}{2}\frac{\sigma^2-1}{\sigma^2}\; x^T x\right),$$

where the superscript T denotes a transpose, so that

$$E\left[w(x)\right] = 1, \quad E\left[w(x)^2\right] = \left(2\sigma^2-1\right)^{-\frac{m}{2}},$$

which gives

$$R = \left(\frac{\sigma^2}{\sqrt{2\sigma^2-1}}\right)^m.$$

The number of samples required for a given M_{eff} thus grows exponentially with the number of variables m:

$$\log(M) = m\ \log\left(\frac{\sigma^2}{\sqrt{2\sigma^2-1}}\right) + \log(M_{\text{eff}}).$$

The situation is illustrated in figure 2. As the number of variables increases, the cost of sampling, measured by the number of samples required to obtain the pre-specified number of effective samples, grows exponentially with the number of variables for any $\sigma > 1/2$ except $\sigma = 1$; see also the discussion in [35].

3 Implicit sampling

Implicit sampling is a general numerical method for constructing effective importance functions [3,12,13,32]. We write the pdf to sample as $f = e^{-F(x)}$ and, temporarily and for simplicity, assume

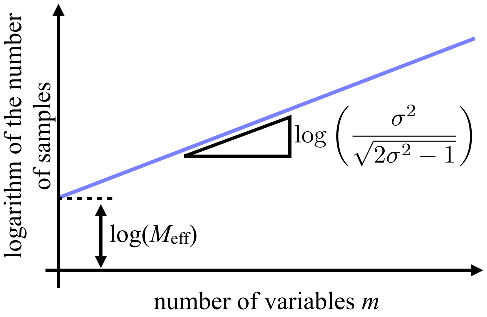

Figure 2: The number of samples required increases exponentially with the number of variables.

that $F(x)$ is convex. The idea is to write x as a one-to-one and onto function of an easy-to-sample reference variable ξ with pdf $g \propto e^{-G(\xi)}$, were G is convex. To find this function, we first find the minima of F, G, call them $\phi_F = \min F, \phi_G = \min G$, and define $\phi = \phi_F - \phi_G$. Then we define a mapping $\psi : \xi \to x$ such that ξ and x satisfy the equation

$$F(x) - \phi = G(\xi). \tag{2}$$

With our convexity assumptions a one-to-one and onto map exists, in fact there are many, because equation (2) is underdetermined (it is a single equation that connects the m elements of X_i to the m elements of Ξ_i, where m is the number of variables). To find samples $X_1, X_2, \dots$, first find samples $\Xi_1, \Xi_2, \dots$ of ξ, and for each $i = 1, 2, \dots$, solve (2). A sample X_i obtained this way has a high probability (with a high probability) for the following reasons. A sample Ξ_i of the reference density g is likely to lie near the minimum of G, so that the difference $(G(\Xi_i) - \phi_G)$ is likely to be small. By equation (2) the difference $(F(X_i) - \phi_F)$ is also likely to be small and the sample X_i is in the region where f is large. Thus, by solving (2), we map the high-probability region of the reference variable ξ to the high-probability region of x, so that one obtains a high-probability sample X_i from a high-probability sample Ξ_i.

The mapping ψ defines the importance function implicitly by the solutions of (2):

$$f_0(x(\xi)) = g(\xi) \left| \det \left(\frac{\partial \xi}{\partial x} \right) \right|.$$

A short calculation shows that the sampling weight $w(X_i)$ is

$$w(X_i) \propto e^{-\phi} \left| J(X_i) \right|, \tag{3}$$

where J is the Jacobian $\det(\partial x / \partial \xi)$. The factor $e^{-\phi}$, common to all the weights, is immaterial here, but plays an important role further below.

We now give an example of a map ψ that makes this construction easy to implement. We assume here and for the rest of the paper that the reference variable ξ is the Gaussian $\mathcal{N}(0, I_m)$, where m is the dimension of x. With a Gaussian ξ, $\phi_G = 0$ and equation (2) becomes

$$F(x) - \phi = \frac{1}{2} \xi^T \xi. \tag{4}$$

We can find a map that satisfies this equation by looking for solutions in a random direction $\eta = \xi / (\xi^T \xi)$, i.e. use a mapping ψ such that

$$x = \mu + \lambda L \eta, \tag{5}$$

where $\mu = \operatorname{argmin} F$ is the minimizer of F, L is a given matrix, and λ is a scalar that depends on ξ. Substitution of the above mapping into (4) gives a scalar equation in the single variable λ (regardless of the dimension of x). The matrix L can be chosen to optimize the distribution of the sampling weights. For example, if L is a square root of the inverse of the Hessian of F at the minimum, then the algorithm is affine invariant, which is important in applications with multiple scales [21]. Equation (5) can be readily solved and the resulting Jacobian is easy to calculate (see [32] for details).

Other constructions of suitable maps ψ are presented in e.g. [12, 20] and some of these are analyzed in [20]. With these constructions, often the most expensive part of the calculation is finding the minimum of F. Note that this needs to be done only once for each sampling problem, and that finding the maximum of f is an unavoidable part of any useful choice of importance function, explicitly or implicitly. Addressing the need for the optimization explicitly has the advantage that effective optimization methods can be brought to bear. Connections between optimization and (implicit) sampling also have been pointed out in [3, 33]. In fact, an alternate derivation of implicit sampling can be given by starting from maximum likelihood estimation, followed by a randomization that removes bias and provides error estimates.

We now relax the assumption that F is convex. If F is U-shaped, the above construction works without modification. A scalar function F is U-shaped if it is piecewise differentiable, its first derivative vanishes at a single point which is a minimum, F is strictly decreasing on one side of the minimum and strictly increasing on the other, and $F(x) \to \infty$ as $|x| \to \infty$; in the general case, F is U-shaped if it has a single minimum and each intersection of the graph of the function $y = F(x)$ with a vertical plane through the minimum is U-shaped in the scalar sense. If F is not U-shaped, but has only one minimum, one can replace it by a U-shaped approximation, say F_0, and then apply implicit sampling as above. The bias created by this approximation can be annulled by reweighting [12]. If there are multiple minima, one can represent f as a mixture of components whose logarithms are U-shaped, and then pick as a reference pdf g a cross-product of a Gaussian and a discrete random variable. However, the decomposition into a suitable mixture can be laborious (see [49]).

4 Beyond implicit sampling

Implicit sampling produces a sequence of samples that lie in the high-probability domain and are weighted by a Jacobian. It is natural to wonder whether one can absorb the Jacobian into the map $\psi : \xi \to x$ and obtain "optimal" samples that need no weights (here and below, optimal refers to minimum variance of the weights, i.e. an optimal sampler is one with a constant weight function). If a random variable x with pdf f is a one-to-one function of a variable ξ with pdf g, then

$$f(x(\xi)) = g(\xi)J(\xi), \tag{6}$$

where J is the Jacobian of the map ψ. One can obtain samples of x directly (i.e. without weights) by solving (6). Notable efforts in that direction can be found in [34, 38].

However, optimal samplers can be expensive to use; the difficulties can be demonstrated by the following construction. Consider a one-variable pdf f with a convex $F = -\log f$. Find the maximizer z of f, with maximum $M_f = f(z)$, and let g be the pdf of a reference variable ξ, with its maximizer also at z and maximum $M_g = g(z)$. To simplify the notations, change variables so that $z = 0$. Then construct a mapping $\psi : \xi \to x$ as follows. Solve the differential equation

$$\frac{du}{dt} = \frac{g(t)}{f(u)}, \tag{7}$$

where t is the independent variable and u is a function of t, in the t-interval $[0, \xi]$, with initial condition $u = 0$ when $t = 0$. Define the map from ξ to x by $x = u(\xi)$; then $f(x) = g(\xi)J$, where $J = |du/dx|$ evaluated at $t = \xi$. It is easy to see that under the assumptions made, this map is one-to-one and onto. We have achieved optimal sampling.

The catch is that the initial value of $g(t)/f(u)$ is M_g/M_f, which requires that the normalization constants of f and g be known. In general the solution of the differential equation does not depend linearly on the initial value of $g(0)/f(u(0))$, so that unless M_f and M_g are known, the samples are in the wrong place while weights to remove the resulting bias are unavailable. Analogous conclusions were reached in [16] for a different sampling scheme. In the special case where $F = -\log f$ and $G = -\log g$ are both quadratic functions the constants M_f, M_g can be easily calculated, but under these conditions one can find a linear map that satisfies equation (6) and implicit sampling becomes identical to optimal sampling. The elimination of all variability in the weights in the nonlinear case is rarely worth the cost of computing the normalization constants.

Note that the normalization constants are not needed for implicit sampling. The mapping (5) for solving (2) is well-defined even when the normalization constants of f and g are unknown. The reason is that if one multiplies the functions f or g by a constant, the logarithm of this constant is added both to F (G) and to ϕ_F (ϕ_G) and cancels out. Implicit sampling simplifies equation (6) by assuming that the Jacobian is a constant. It produces weighted samples where the weights are known only up to a constant, and this indefiniteness can be removed by normalizing the weights so that their sum equals 1.

5 Bayesian parameter estimation

We now apply implicit sampling to Bayesian estimation. For the sake of clarity we first explain how to use Bayesian estimation to determine physical parameters needed in the numerical modeling of physical processes, given noisy observations of these processes. This discussion explains how a prior and data combine to create a posterior proability density that is used for inference. In the next section we extend the methodology to data assimilation, which is our main focus.

Suppose one wishes to model the diffusion of some material in a given medium. The density of the diffusing material can be described by a diffusion equation, provided the diffusion coefficients $\theta \in \mathbb{R}^m$ can be determined. Given the coefficients θ, one can solve the equation and determine the values of the density at a set of points in space and time. The values of the density at these points can be measured experimentally, by some method with an error η whose probability density is assumed known; once the measurements $d \in \mathbb{R}^k$ are made, this assigns a probability to θ. The relation between d and θ can be written as

$$d = h(\theta) + \eta, \tag{8}$$

where $h : \mathbb{R}^k \to \mathbb{R}^m$ is a generally nonlinear function, and $\eta \sim p_\eta(\cdot)$ is a random variable with known pdf that represents the uncertainty in the measurements. The evaluation of h often requires a solution of a differential equation and can be expensive. In the Bayesian approach, one assumes that information about the parameters is available in form of a prior $p_0(\theta)$. This prior and the likelihood $p(d|\theta) = p_\eta(d - h(\theta))$, defined by (8), are combined via Bayes' rule to give the posterior pdf, which is the probability density function for the parameters θ given the data d,

$$p(\theta|d) = \frac{1}{\gamma(d)} p_0(\theta) p(d|\theta), \tag{9}$$

where $\gamma(d) = \int p_0(\theta) p(d|\theta) d\theta$ is a normalization constant (which is hard to compute).

In a Monte Carlo approach to the determination of the parameters, one finds samples of the posterior pdf. These samples can, for example, be used to approximate the expected value

$$E[\theta|d] = \int \theta \, p(\theta|d) \, d\theta,$$

which is the best approximation of θ in a least squares sense (see, e.g. [8]), and a error estimate can also be computed, see also [43].

When sampling the posterior $p(\theta|d)$, it is natural to set the importance function equal to the prior probability $p_0(\theta)$ and then weight the samples by the likelihood $p(d|\theta)$. However, if the data are informative, i.e., if the measurements d modify the prior significantly, then the prior and the posterior can be very different and this importance sampling scheme is ineffective. When implicit sampling is used to sample the posterior, then the data are taken into account already when generating the samples, and not only in the weights of the samples.

As the number of variables increases, the prior p_0 and the posterior $p(\theta|d)$ may become nearly mutually singular. In fact, in most practical problems these pdfs are negligible outside a small spherical domain so that the odds that the prior and the posterior have a significant overlap decrease quickly with the the number of variables, making implicit sampling increasingly useful (see [33] for an application of implicit sampling to parameter estimation).

6 Data assimilation

We now apply Bayesian estimation via implicit sampling to data assimilation, in which one makes inferences from an unreliable time-dependent approximate model that defines a prior, supplemented by a stream of noisy and/or partial data. We assume here that the model is a discrete recursion of the form

$$x^{n+1} = f(x^n) + w^n, \tag{10}$$

where n is a discrete time, $n = 0, 1, 2, \dots$, the set of m-dimensional vectors $x^{0:n} = (x^0, x^1, \dots, x^n, \dots)$ are the state vectors we wish to estimate, $f(\cdot)$ is a given function, and w^n is a Gaussian random vector $\mathcal{N}(0, Q)$. The model often represents a discretization of a stochastic differential equation. We assume that k-component data vectors b_n, $n = 1, 2, \dots$ are related to the states x^n by

$$b^n = h(x^n) + \eta^n, \tag{11}$$

where $h(\cdot)$ is the (generally nonlinear) observation function and η^n is a $\mathcal{N}(0, R)$ Gaussian vector (the η^n and w^n are independent of η^k and w^k for $k < n$ and also independent of each other). In addition, initial data x^0, which may be random, are given at time $n = 0$. For simplicity, we consider in this section the problem of estimating the state x of the system rather than estimating parameters in the equations as in the previous section. Thus, we wish to sample the conditional pdf $p(x^{0:n}|b^{1:n})$, which describes the probability of the sequence of states between 0 and n given the data $b^{1:n} = (b^1, \dots, b^n)$. The samples of this conditional pdf are sequences $X^0, X^1, X^2, \dots$, usually referred to as “particles”. The conditional pdf satisfies the recursion

$$p(x^{0:n+1}|b^{1:n+1}) = p(x^{0:n}|b^{1:n}) \frac{p(x^{n+1}|x^n) p(b^{n+1}|x^{n+1})}{p(b^{n+1}|b^{1:n})}, \tag{12}$$

where $p(x^{n+1}|x^n)$ is determined by equation (10) and $p(b^{n+1}|x^{n+1})$ is determined by equation (11). (see e.g. [15]). We wish to sample the conditional pdf recursively, time step after time step, which is natural in problems where the data are obtained sequentially. To do this we use an importance function π of the form

$$\pi(x^{0:n+1}|b^{1:n+1}) = \pi_0(x^0) \prod_{k=1}^{n+1} \pi_k(x^k|x^{0:k-1}, b^{1:k}). \tag{13}$$

where the π_k are increments of the importance function. The recursion (12) and the factorization (13) yield a recursion for the weight W_j^{n+1} of the j-th particle,

$$W_j^{n+1} = \frac{p(x^{0:n+1}|b^{1:n+1})}{\pi(x^{0:n+1}|b^{1:n+1})} = W_j^n \frac{p(X_j^{n+1}|X_j^n)p(b^{n+1}|X_j^{n+1})}{\pi_{n+1}(X_j^{n+1}|X^{0,n}, b^{1:n})}. \tag{14}$$

With this recursion, the task at hand is to pick a suitable update $\pi_k(x^k|x^{0:k-1}, b^{1:k})$ for each particle, to sample $p(X_j^{n+1}|X_j^n)p(b^{n+1}|X_j^{n+1})$, and to update the weights so that the pdf $p(x^{0:n+1}|b^{1:n+1})$ is well-represented. Thus, the solution of the discrete model (10) plays the same role in data assimilation as the prior in the parameter estimation problem of section 5.

In the SIR (sequential importance sampling) filter one picks

$$\pi_{n+1}(x^{n+1}|x^{0:n}, b^{1:n+1}) = p(x_j^{n+1}|X_j^{n-1}),$$

and the weight is $W_j^{n+1} \propto W_j^n \, p(b^{n+1}|X_j^{n+1})$. Thus, the SIR filter proposes samples by the model (10) and determines their probability from the data (11).

In implicit data assimilation one samples $\pi_{n+1}(x^{n+1}|x^{0:n}, b^{1:n+1})$ by implicit sampling, thus taking the most recent data b^{n+1} into account for generating the samples. The weight of each particle is given by equation (14)

$$W_j^{n+1} = W_j^n \exp(-\phi_j^{n+1}) J(X_j^{n+1}),$$

where $\phi_j^{n+1} = \text{argmin } F_j^{n+1}$ is the minimum of

$$F_j^{n+1}(x^{n+1}) = -\log\left(p(x^{n+1}|X_j^n)p(b^{n+1}|x^{n+1})\right).$$

Thus, a minimization is required for each particle at each step.

The "optimal" importance function [16, 26, 51] is defined by (13) with

$$\pi_{n+1}(x^{n+1}|x^{0:n}, b^{1:n+1}) = p(x^{n+1}|x^n, b^{n+1}),$$

and its weights are

$$W_j^{n+1} = W_j^n \; p(b^{n+1}|X_j^n).$$

This choice of importance functions is optimal in the sense that it minimizes the variance of the weights at the $n+1$ step given the data and the particle positions at the previous step. In fact, this importance function is optimal over all importance functions that factorize as in (13), and in the sense that the variance of the weights $\text{var}(w^n)$ is minimized (with expectations computed with respect to $\pi(x^{0:n+1}|b^{0:n+1})$), see [42]. In a linear problem where all the noises are Gaussian, the implicit filter and the optimal filter coincide [13, 29, 32]. When a problem is nonlinear, the optimal filter may be hard to implement, as discussed above.

A variety of other constructions is available (see e.g. [1, 46–49]). The SIR and optimal filters are the two extreme cases, in the sense that one of them makes no use of the data in finding samples while the other makes maximum use of the data. The optimal filter becomes impractical for nonlinear problems while the implicit filter can be implemented at a reasonable cost. Implicit sampling thus balances the use of data in sampling and the computational cost.

7 The size of a covariance matrix and the feasibility of data assimilation

Particle filters do not necessarily succeed in assimilating data. There are many possible causes- for example, the recursion (10) may be unstable, or the data may be inconsistent. The most frequent cause of failure is filter collapse, in which only one particle has a non-negligible weight, so that there is effectively only one particle, which is not sufficient for valid inferences. In the next section we analyze how particle collapse happens and how to avoid it. In preparation for this discussion, we discuss here the Frobenius norm of a covariance matrix, its geometrical significance, and physical interpretation.

The effectiveness of a sampling algorithm depends on the pdf that it samples. For example, suppose that the posterior you want to sample is in one variable, but has a large variance. Then there is a large uncertainty in the state even after the data are taken into account, so that not much can be said about the state. We wish to assess the conditions under which the posterior can be used to draw reliable conclusions about the state, i.e. we make the statement "the uncertainty is not too large" quantitative. We call sampling problems with a small enough uncertainty "feasible in principle". There is no reason to attempt to solve a sampling problem that is not feasible in principle.

Our analysis is linear and Gaussian and we allow the number of variables to be large. in multivariate Gaussian problems, feasibility requires that a suitable norm of the covariance matrix be small enough. This analysis is inspired by geophysical applications where the number of variables is large, but the nonlinearity may be mild; parts of it were presented in [11].

We describe the size of the uncertainty in multivariate problems by the size the $m \times m$ covariance matrix $P = (p_{ij})$, which we measure by its Frobenius norm

$$||P||_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{m} p_{ij}^2 \right)^{1/2} = \left(\sum_{k=1}^{m} \lambda_k^2 \right)^{1/2}, \tag{15}$$

where the λ_k, $k = 1, \dots, m$ are the eigenvalues of P. The reason for this choice is the geometric significance of the Frobenius norm. Let $x \sim \mathcal{N}(\mu, P)$ be an m-dimensional Gaussian variable, and consider the distance between a sample of x and its most likely value μ

$$r = ||x - \mu||_2 = \sqrt{(x-\mu)^T (x-\mu)}.$$

If all eigenvalues of P are $O(1)$ and if m is large, then $E[r]$, the expected value of r, is $O(||P||_F)$ and $var(r) = O(1)$, i.e. the samples concentrate on a thin shell, far from their most likely value. Different parts of the shell may correspond to physically different scenarios (such as "rain" or "no rain"). Moreover, the expected value and variance of the distance distance $r = \sqrt{y}$ of a sample to its most likely value are bounded above and below by multiples of $||P||_F$ (see [11]). If one tries to estimate the state of a system with a large $||P||_F$, the Euclidean distance between a sample and the resulting estimate is likely to be large, making the estimate useless. For a feasible problem we thus require that $||P||_F$ not be too large. How large "large" can be depends on the problem.

In [11] and in earlier work [4, 6, 41] on the feasibility of particle filters, the Frobenius norm of P was called the "effective dimension". This terminology is confusing and we abandon it here. We show below that $||P||_F$ quantifies the strength of the noise. We define the effective dimension of the noise by

$$\min \ell \in \mathbb{Z} : \sum_{j=1}^{\ell} \lambda_j^2 \geq (1-\epsilon) \sum_{j=1}^{\infty} \lambda_j^2, \tag{16}$$

where ϵ is a predetermined small parameter. A small $||P||_F$ can imply a small ℓ, however the reverse is not necessarily true and $||P||_F$ can be small, even though ℓ is large (for example if $x \sim \mathcal{N}(0, I_m/\sqrt{m})$) and vice versa. There are small dimensional problems with a large noise (large $||P||_F$) which are not feasible in principle, and there are large dimensional problems with a small noise which are feasible in principle (small $||P||_F$). Estimation based on a posterior pdf is feasible in principle only if $||P||_F$ is small.

We now examine the relations between $||P||_F$, the effective dimension ℓ, and the correlations between the components of the noise (i.e., the size of the off-diagonal terms in P). We assume that our random variables are point values x_i of a stationary stochastic process on the real line, measured at the points ih of the finite interval $[0, 1]$, where $i = 1, 2, \ldots, m$ and $h = 1/m$. As the mesh is refined, m increases. The condition for the discrete problem to be feasible is that the Frobenius norm of its covariance matrix be bounded. Let the covariance matrix of the continuous process be $k(x, x') = k(x - x')$; the corresponding covariance operator is defined by

$$K\varphi = \int_0^1 k(x, x')\varphi(x')dx'$$

for every function $\varphi = \varphi(x)$. The eigenvalues of K can be approximated by the eigenvalues of P multiplied by h (see [37]). The Frobenius norm of K,

$$\int_0^1 \int_0^1 k^2 \, dx \, dx',$$

describes the distance of a sample to its most likely value in the L^2 sense, as can be seen by following the same steps as in [11], replacing the Euclidean norm by the appropriate grid-norm, i.e. $||x||_2^2 = \sum_{k=1}^m h \, x_k^2$.

We consider a family of Gaussian stationary processes with differing correlation structures. The discussion is simplified if we keep the energy e defined by

$$e = \int_{-\infty}^{\infty} k(y, y')^2 \, dy'.$$

equal to 1 for all the members of the family (so that all the members of the family are equally noisy); that e is an energy follows from Khinchin's theorem (see e.g. [8]). An example of such a family is the family of zero-mean processes with

$$k(x, x') = \pi^{-\frac{1}{4}} L^{-\frac{1}{2}} \exp\left(-\frac{(x - x')^2}{2L^2}\right),$$

where L is the correlation length. The infinite limits in the definition of e make the calculations easier, and are harmless as long as L is less than about $1/2$. The elements p_{ij} of the discrete matrix P are $p_{ij} = k(ih, jh)$, $i, j = 1, \ldots, m$.

In the left panel of figure 3 we demonstrate that for a fixed correlation length $L = 0.1$, the Frobenius norm

$$||P||_F = \left(\sum_{k=1}^m (h \, \lambda_k)^2\right)^{\frac{1}{2}},$$

where the λ_k are the eigenvalues of the covariance matrix P, is independent of the mesh (for small enough h) and, therefore independent of the discretization. In the calculation of the dimension ℓ in (16), we use $\epsilon = 5\%$. In the figure, we plot the Frobenius norms of the $m \times m$ matrices P

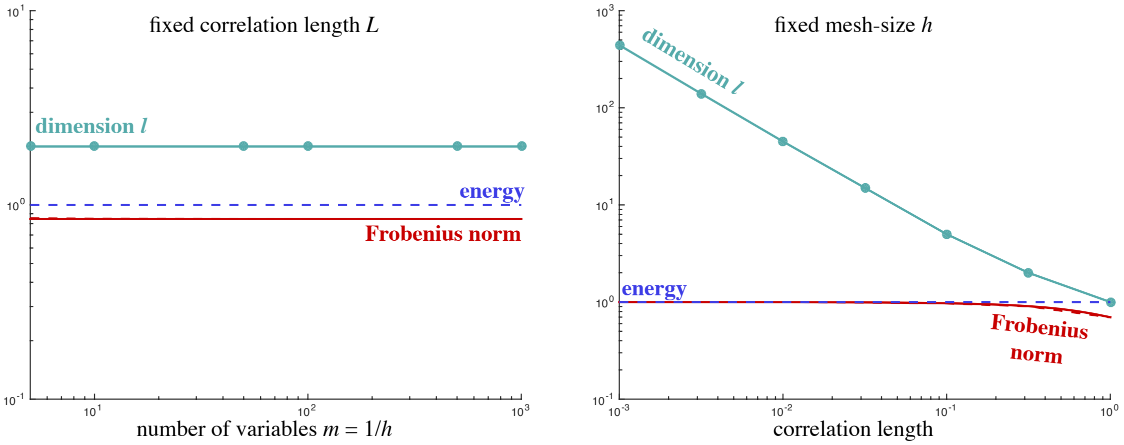

Figure 3: Norms of covariance matrices and dimension ℓ for a family of Gaussian processes with constant energy as a function of the mesh size h (left) and as a function of the correlation length L (right).

as the purple and red) dashed lines. The solid line is the Frobenius of the covariance operator K of the continuous process, which, for this simple example, can be computed analytically from the expansion

$$k(y, y') = \sum_{j=1}^{\infty} \tilde{\lambda}_j e_j(y) e_j(y'),$$

where $\tilde{\lambda}$ is an eigenvalue of $k(y, y')$ with eigenvector $e(y)$. The eigenvalues and eigenvectors are defined by

$$\int_0^1 k(y, y')\, e(y)\; dy = \tilde{\lambda}\, e(y'),$$

and

$$\int_0^1 e_i(y) e_j(y)\; dy = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Thus,

$$\int_0^1 \int_0^1 k(y, y)^2\; dydy' = \sum_{j=1}^{\infty} \tilde{\lambda}_j^2 = \pi\, L \left(\left(\exp\left(-\frac{1}{L^2} \right) - 1 \right) + \sqrt{\pi}\, \mathrm{Erf}\left(\frac{1}{L} \right) \right)$$

Since $\tilde{\lambda} \approx h\lambda$, we have that

$$||P||_F^2 \approx \sum_{j=1}^{\infty} \tilde{\lambda}_j^2 = \pi\, \left(L \left(\exp\left(-\frac{1}{L^2} \right) - 1 \right) + \sqrt{\pi}\, \mathrm{Erf}\left(\frac{1}{L} \right) \right).$$

We find good agreement between the infinite dimensional results and their finite dimensional approximations (the dashed lines are mostly invisible because they coincide with the results calculated from the infinite dimensional problem).

The right panel of the figure shows the variation of $||P||_F$ and the dimension ℓ with the correlation length L. We observe that the Frobenius norm remains almost constant for $L < 10^{-2}$. On the other hand, the dimension ℓ in (16) increases as the correlation length decreases. What this figure shows is that the feasibility of data assimilation depends on the level of noise, but not on the effective number of variables. One can find the same level of noise for very different effective dimensions. If data assimilation is not feasible, it is because the problem has too much noise, not too many variables.

It is interesting to consider the limit $L \to 0$ in our family of processes. A stationary process $u(x)$ such that for every x, $u(x)$ is a Gaussian variable, with $E[u(x)u(x')] = 0$ for $x \neq x'$ and $E[u(x)^2] = A$, where A is a finite constant, has very little energy; white noise, the most widely used Gaussian process with independent values, has $E[u(x)u(x')] = \delta_{x,x'}$, where the right-hand-side is a δ function, so that $E[u(x)^2]$ is unbounded. The energy of white noise is infinite, as is well-known. If $k(x-x')$ in our family blew up like L^{-1} at the origin as $L \to 0$, the process would be a multiple of Brownian motion; here it blows up like $L^{-1/2}$, allowing the energy to remain finite while not allowing the $E[u^2]$ to remain bounded. The moral of this discussion is that a sampling problem where $P = I_m$ for all m is unphysical.

An apparent paradox appears when one applies our theory to determine the feasibility of a Gaussian random variable with covariance matrix $P = I_m$, as in [4,6,40–42]. In this case, $||P||_F = \sqrt{m}$, so that the problem is infeasible by our definition if the number of variables m is large. On the other hand, the problem seems to be physically plausible. For example, suppose one wishes to estimate the wind velocity at m geographical sites based on independent local models for each site (e.g. today's wind velocity is yesterday's wind velocity with some uncertainty), and independent noisy measurements at each site. The local P_j at each site is one-dimensional and equals 1. Why then not try to determine the m velocities in one fell swoop, by considering the whole vector x and setting $P = I_m$? Intuition suggest that the set of local problems is equivalent to the $P = I_m$ problem, e.g. that one can use local stochastic models to predict the velocities at nearby sites, and then mark these on a weather map and obtain a plausible global field. Thus, $||P||_F$ is large in a plausible and feasible problem. This, however, is not so. First, in reality, the uncertainties in the velocities at nearby sites are highly correlated, while the problem $P = I_m$ assumes that this is not so. The resulting velocities map from the latter will be unphysical and lead to wrong forecasts- large scale coherent flows in today's weather map will have a different impact on tomorrow's forecast than a set of uncorrelated wind patterns, and of course carry a much larger energy, even if the local amplitudes are the same. If one has a global problem where the covariance matrix is truly I_m, then replacing the solution of the full problem by a component-wise solution changes estimate of the noise from $||P||_F$ to the numerically smaller maximum of the component variances, which is not as good a measure of the distance between the samples and their mean.

8 Linear analysis of the convergence of data assimilation

We now summarize the linear analysis of the conditions under which particular particle filters produce reliable estimates when the problem is linear (see [11]). A general nonlinear analysis is not within reach, while the linear analysis captures the main issues.

The dynamical equation now takes the form:

$$x^{n+1} = Ax^n + w^n, \tag{17}$$

where A is an $m \times m$ matrix and, the data equation becomes

$$b^{n+1} = Hx^{n+1} + \eta^n, \tag{18}$$

where H is an $k \times m$ matrix. As before, w^n are independent $\mathcal{N}(0, Q)$ and the η^n are independent $\mathcal{N}(0, R)$, independent also of the w^n. We assume that equation (17) is stable and the data are sufficient for assimilation (for a more technical discussion of these assumptions, see e.g. [11]). These two equations define a posterior probability density, independently of any method of estimation. The first question is, under what conditions is state estimation with the posterior feasible in principle. If one wants to estimate the state at time n, given the data up to time n, this requires that the covariance of $x^n|b^{1:n}$ have a small Frobenius norm.

The theoretical analysis of the Kalman filter [23] can be used to estimate this covariance, even when the problem is not feasible in principle or the Kalman filter itself is too expensive for practical use. Under wide conditions, the covariance of $x^n|b^{1:n}$ rapidly reaches a steady state

$$P = (I - KH)X,$$

where X is the unique positive semi-definite solution of a particular nonlinear algebraic Riccati equation (see e.g. [25]) and

$$K = XH^T(HXH^T + R)^{-1}.$$

One can calculate $||P||_F$ and decide whether the estimation problem is feasible in principle. Thus, a condition for successful data assimilation is that $||P||_F$ be moderate, which generaly requires that the $||Q||_F$, $||R||_F$ in equations (17) and (18) be small enough.

A particle filter can be used to estimate the state by sampling the posterior pdf $p(x^{0:n+1}|b^{1:n+1})$, defined by (17) and (18). The state at time n conditioned on the data up to time n can be computed by marginalizing samples of $p(x^{0:n+1}|b^{1:n+1})$ to samples of $p(x^n|b^{1:n+1})$ (which amounts to simply dropping the sample's past). Thus, a particle filter does not directly sample $p(x^n|b^{1:n+1})$, so that its samples carry weights (even in linear problems). These weights must not vary too much or else the filter "collapses" (see section 2). Thus, a particle filter can fail, even if the estimation problem is feasible in principle.

It was shown in [4, 6, 11, 40, 41] that the variance of the negative logarithm of the weight must be small or else the filter collapses. For the SIR filter, the condition that this collapse not happen is that the Frobenius norm of the matrix

$$\Sigma_{\text{SIR}} = H(Q + APA^T)H^TR^{-1}, \tag{19}$$

be small enough. For the optimal filter, which in the present linear setting coincides with the implicit filter, success requires a small Frobenius norm for

$$\Sigma_{\text{Opt}} = HAPA^TH^T(HQH^T + R)^{-1}, \tag{20}$$

see [11]. In either case, this additional condition must be satisfied as well as the the condition that $||P||_F$ is small.

To understand what these formulas say, we apply them to a simple model problem which is popular as a test problem in the analysis of numerical weather prediction schemes. The problem is defined by $H = A = I_m$ and $Q = qI_m$, $R = rI_m$, where we vary the number of components m. Note that for a fixed m, the problem is parameterized by the variance r of the noise in the model and the variance q of the noise in the data. This problem is feasible in principle if

$$||P||_F = \sqrt{m}\,\frac{\sqrt{q^2 + 4qr} - q}{2},$$

is not too large. For a fixed m, this means that

$$\frac{\sqrt{q^2 + 4qr} - q}{2} \leq 1,$$

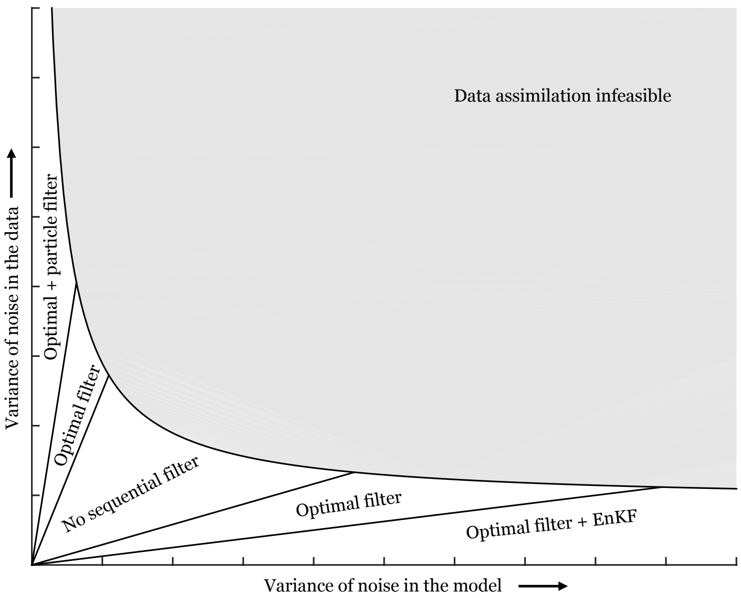

Figure 4: Conditions for feasible data assimilation and for successful sampling with the optimal particle filter, the particle filter and the ensemble Kalman filter (from [31]).

where the number 1 stands in for a sharper estimate of the acceptable variance for a given problem (and this choice gives the complete qualitative story). In figure 4, this condition is illustrated and shown are all feasible problems in white and all infeasible problems in grey. The analysis thus shows that for data assimilation to be feasible, either the model or the data must be accurate. If both are noisy, the noise dominates so that estimation is useless.

In the SIR filter, the additional condition for success becomes

$$\frac{\sqrt{q^2+4qr}+q}{2r} \leq 1,$$

and for the optimal/implicit case, the additional condition is

$$\frac{\sqrt{q^2+4qr}-q}{2(q+r)} \leq 1. \tag{21}$$

In both cases, the number 1 on the right-hand-side stands in for a sharper estimate of the acceptable variance of the weights for a given problem (which also depends on the computing resources). In either case, the added condition is quadratic and homeogeneous in the ratio q/r, and thus slices out conical regions from the region where data assimilation is feasible in principle. These conical regions are shown in figure 4. The region where estimation with a particular filter is feasible is labeled with the name of the particular filter. Note that we also show results for the ensemble Kalman filter (EnKF) [17, 44], which are derived in [31], but not discussed in the present paper. The analysis of the EnKF relates to the situation where it is used to sample the same posterior density as the other filters quoted in the figure. The analysis in [31] also includes the situation where the EnKF is used to sample a marginal of that pdf directly, when the conclusions are different.

In practical problems one usually tries to sharpen estimates obtained from approximate dynamics with the help of accurate measurements, and the optimal/implicit filter works well in such problems. However, there is a region in which not even the optimal filter succeeds. What fails there is the sequential approach to the estimation problem, i.e. the factorization of the importance function as in (13), see [31, 42]. Non-recursive filters can succeed in this region, and there too implicit sampling can be helpful i.e. one can apply implicit sampling direction to $p(x^{0:n}|b^{1:n})$.

One conclusion we reach from our analysis is that if Bayesian estimation is feasible in principle then it is feasible in practice. However, it is assumed in the previous analysis so far that one has suitable priors, or equivalently, one knows what the noise w^n in equation (10) really is. The fact is that generally one does not. We discuss this issue in the next section.

9 Estimating the prior

The previous sections described how one may use a prior and a distribution of observation errors to estimate the states or the parameters of a system. This estimation depends on knowing the prior. As we have seen, in data assimilation, the prior is generated by the solution of an equation such as the recursion (10) and depends on knowing the noise w^n in that equation. It is often assumed that the noise in that equation is white. However, one can show that the noise is not white in most problems (see e.g. [9, 14, 50]). We now present a preliminary discussion of methods for estimating the noise.

First, one has to make some assumptions about the origin of the noise. A reasonable assumption (see e.g. [5, 9]) is that there is noise in the equations of motion because a practical description of nature is necessarily incomplete. For example, one can write a solvable equation of motion for a satellite turning around the earth by assuming that the gravitational pull is due to a perfectly spherical earth with a density that is a function only of distance from the center (see [30]). Reality is different, and the difference produces noise, also known as model error. The problems to be solved are: (i) estimate this difference, (ii) identify it, i.e., find a concise approximate representation of this difference that can be effectively evaluated or sampled on a computer, and (iii) design an algorithm that imbeds the identified noise in a practical data assimilation scheme. These problems have been discussed in a number of recent papers, e.g. [9, 28], in particular the review paper [22] which contains an extensive bibliography.

Assume that the full description of a physical system has the form:

$$\frac{d}{dt}x = R(x, y), \quad \frac{d}{dt}y = S(x, y), \tag{22}$$

where t is the time, $x = (x_1, x_2, \ldots, x_m)$ is the vector of resolved variables, and $y = (y_1, y_2, \ldots, y_\ell)$ is the vector of unresolved variables, with initial data $x(0), y(0)$. Consider a situation where this system is too complicated to solve, but where data are available, usually as sequences of measured values of x, assumed here to be observed with negligible observation errors. Write $R(x, y)$ in the form

$$R(x, y) = R_0(x) + z(x, y), \tag{23}$$

where R_0 is a practical approximation of $R(x, y)$ and one is able to solve the equation

$$\frac{d}{dt}x = R_0(x). \tag{24}$$

However, x does not satisfy equation (24) because the true equation, the first of equations (22), is more complicated. The difference between the true equation and the approximate equation is the remainder $z(x, y) = R(x, y) - R_0(x)$, which is the noise in the determination of x. In general, z has to be determined from the data, i.e., from the observed values of x.

A usual approach to the problem of estimating z is to use equation (23) to obtain its values from x data, i.e. calculate $z = \frac{d}{dt}x - R_0(x)$, and then identify it as a continuous stochastic process. This may be difficult, in particular, calculating z requires that one differentiate x, which is generally impractical or inaccurate because z may have high-frequency components or fail to be sufficiently smooth, and the data may not be available at sufficiently small time intervals (an illuminating analysis in a special case can be found in [36, 39]). Once one has values of z, identifying it as a function of the continuous variable t often requires making unwarranted assumptions on the small-scale structure of z, which may be unknown when the data are available only at discrete times.

An alternative is supplied by a discrete approach as in [10]. Equation (24) is always solved on the computer, i.e., in discrete form, the data are always given at discrete points, and it is x one wishes to determine but in general one is not interested in determining z per se. We can therefore avoid the difficult detour through a continuous-time z followed by a discretization, as follows. We pick once and for all a particular discretization of equation (24) with a particular time step time step δ,

$$x^{n+1} = x^n + \delta R_\delta(x^n), \tag{25}$$

where R_δ is obtained, for example, from a Runge–Kutta method, and where n indexes the result after n steps; the differential equation has been reduced to a recursion such as equation (10). We then use the data to identify the discrepancy sequence, $z_\delta^{n+1} = (x^{n+1} - x^n)/\delta - R_\delta(x^n)$, which is available from x data without approximation.

Then assume, as one does in the continuous-time case, that the system under consideration is ergodic, so that its long-time statistics are stationary. The sequence z_δ^n becomes a stationary time series, which can be represented by one of the representations of time series, e.g. the NARMAX (nonlinear auto-regression moving average with exogenous inputs) representation, with x as an exogenous input. The observed x of course may include observation errors, which have to be separated from the model noise via some version of filtering.

As an example, we applied this construction to the Lorenz 96 system [27], created to serve as a metaphor for the atmosphere, which has been widely used as a test bench for various reduction and filtering schemes. It consists of a set of chaotic differential equations, which, following [18], we write as:

$$\begin{aligned} \frac{d}{dt}x_k \quad &= x_{k-1}\left(x_{k+1} - x_{k-2}\right) - x_k + F + z_k, \\ \frac{d}{dt}y_{j,k} \quad &= \tfrac{1}{\varepsilon}[y_{j+1,k}(y_{j-1,k} - y_{j+2,k}) - y_{j,k} + h_y x_k] \end{aligned} \tag{26}$$

with

$$z_k = \frac{h_x}{J}\sum_j y_{j,k},$$

and $k = 1, \ldots, K$, $j = 1, \ldots, J$. The indices are cyclic, $x_k = x_{k+K}$, $y_{j,k} = y_{j,k+K}$ and $y_{j+J,k} = y_{j,k+1}$. This system is invariant under spatial translations, and the statistical properties are identical for all x_k. The parameter ε measures the time-scale separation between the resolved variables x_k and the unresolved variables $y_{j,k}$. The parameters are $\varepsilon = 0.5$, $K = 18, J = 20, F = 10, h_x = -1$ and $h_y = 1$. The ergodicity of the Lorenz 96 system has been established numerically in earlier work (see e.g. [18]). We pretend that this system is too difficult to integrate in full; we take as as R_0 of equation (24) the system is which is $z_k = 0$. The noise is then z_k, which in our special case can actually be calculated by solving the full system of equations; in general the noise has to be estimated from data as described above and in [10, 50]. In figure 5 we plot the covariance function

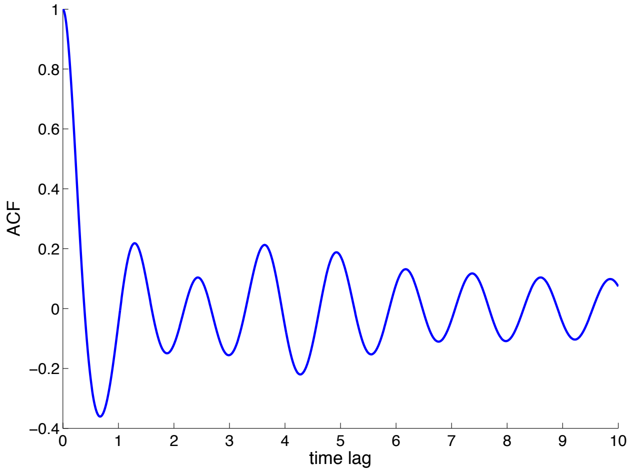

Figure 5: Autocorrelation function of the noise z in the Lorenz 96 system.

of the noise component z_1 determined as just described. Note that z cannot be thought of white noise even remotely. To the extent that the Lorenz 96 is a valid metaphor for the atmosphere, we find that the noise in a realistic problem is not white. This can of course be deduced from the construction after equations (22), where the noise appears as the difference between solutions of differential equations, which is not likely to be white.

This relation between the preceding discussion of the noise (which defines the prior through equation (23)) should be compared with the discussion of implicit sampling and particle filters. Here too the data have been put to additional use (there to define samples and not only to weight them, here to provide information about the noise as well as about the signal); here too an assumption of a white noise input has been found wanting, and realism requires non-trivial correlations, (there in space, here in time).

The next step is to combine the recursion (25) with observations to produce a state estimate. Recent work on this topic is summarized in [22]. An example of this construction could not be produced in time to appear in this article.

10 Conclusions

We have presented algorithms for data assimilation and for estimating the prior, with some analysis. The work presented shows that the keys to success are a better use of the data and a careful analysis of what is physically plausible and useful. We feel that significant progress has been made. One conclusion we draw from this work is that data assimilation, which is often considered as a problem in statistics, should also, or even mainly, be viewed as a problem in computational physics. This conclusion has also been reached by others, see e.g. [14, 28].

11 Acknowledgements

This work was supported in part by the Director, Office of Science, Computational and Technology Research, U.S. Department of Energy, under Contract No. DE-AC02-05CH11231, and by the National Science Foundation under grants DMS-1217065 and DMS-1419044.

References

[1] M. Ades and P.J. van Leeuwen. An exploration of the equivalent weights particle filter. *Quarterly Journal of the Royal Meteorological Society*, 139(672):820–840, 2013.

[2] M.S. Arulampalam, S. Maskell, N.J. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transaction on Signal Processing*, 50(2):174 –188, 2002.

[3] E. Atkins, M. Morzfeld, and A.J. Chorin. Implicit particle methods and their connection with variational data assimilation. *Monthly Weather Review*, 141:1786–1803, 2013.

[4] T. Bengtsson, P. Bickel, and B. Li. Curse of dimensionality revisited: the collapse of importance sampling in very large scale systems. *IMS Collections: Probability and Statistics: Essays in Honor of David A. Freedman*, 2:316–334, 2008.

[5] T. Berry and J. Harlim. Linear theory for filtering nonlinear multiscale systems with model error. *Proc. R. Soc. A*, 470:20140168, 2014.

[6] P. Bickel, T. Bengtsson, and J. Anderson. Sharp failure rates for the bootstrap particle filter in high dimensions. *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, 3:318–329, 2008.

[7] M. Bocquet, C.A. Pires, and L. Wu. Beyond Gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review*, 138:2997–3023, 2010.

[8] A. J. Chorin and O. H. Hald. *Stochastic tools in mathematics and science.* Springer, third edition, 2013.

[9] A.J. Chorin and O.H. Hald. Estimating the uncertainty in underresolved nonlinear dynamics. *Math. Mech. Solids*, 19:28–38, 2014.

[10] A.J. Chorin and F. Lu. A discrete approach to stochastic parametrization and dimensional reduction in nonlinear dynamics. *arXiv:1503.08766*, 2015.

[11] A.J. Chorin and M. Morzfeld. Conditions for successful data assimilation. *Journal of Geophysical Research – Atmospheres*, 118:11,522–11,533, 2013.

[12] A.J. Chorin, M. Morzfeld, and X. Tu. Implicit particle filters for data assimilation. *Communications in Applied Mathematics and Computational Science*, 5(2):221–240, 2010.

[13] A.J. Chorin and X. Tu. Implicit sampling for particle filters. *Proceedings of the National Academy of Sciences*, 106(41):17249–17254, 2009.

[14] D. Crommelin and E. Vanden-Eijnden. Subgrid-scale parameterization with conditional Markov chains. *J. Atmos. Sci*, 65:2661–2675, 2008.

[15] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo methods in practice.* Springer, 2001.

[16] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10:197–208, 2000.

[17] G. Evensen. *Data assimilation: the ensemble Kalman filter.* Springer, 2006.

[18] I. Fatkulin and E. Vanden-Eijnden. A computational strategy for multi-scale systems with applications to Lorenz'96 model. *J. Comput. Phys.*, 200:605–638, 2004.

[19] A. Fournier, G. Hulot, D. Jault, W. Kuang, W. Tangborn, N. Gillet, E. Canet, J. Aubert, and F. Lhuillier. An introduction to data assimilation and predictability in geomagnetism. *Space Science Review*, 155:247–291, 2010.

[20] J. Goodman, K.K. Lin, and M. Morzfeld. Small-noise analysis and symmetrization of implicit monte carlo samplers. *Communications on Pure and Applied Mathematics*, accepted, 2015.

[21] J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Sciences*, 5(1):65–80, 2010.

[22] J. Harlim. Data assimilation with model error from unresolved scales. *arXiv:1311.3579*, 2013.

[23] R.E. Kalman. A new approach to linear filtering and prediction theory. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–48, 1960.

[24] M.H. Kalos and P.A. Whitlock. *Monte Carlo methods*, volume 1. John Wiley & Sons, 1 edition, 1986.

[25] P. Lancaster and L. Rodman. *Algebraic Riccati equations.* Oxford University Press, 1995.

[26] J.S. Liu and R. Chen. Blind deconvolution via sequential imputations. *Journal of the American Statistical Association*, 90(430):567–576, 1995.

[27] E.N. Lorenz. Predictability: A problem partly solved. *Proc. Seminar on predictability*, pages 1–18, 1995.

[28] A.J. Majda and J. Harlim. Physics constrained nonlinear regression models for time series. *Nonlinearity*, 26(1):201–217, 2013.

[29] R. Miller and L. Ehret. Application of the implicit particle filter to a model of nearshore circulation. *Journal of Geophysical Research*, 119:doi: 10.1002/2013JC009440, 2014.

[30] M. Morzfeld and H.C. Godinez. Orbit determination and prediction with Bayesian statistics and implicit sampling. *preprint*, 2015.

[31] M. Morzfeld and D. Hodyss. Analysis of the ensemble kalman filter for marginal and joint posteriors. *Monthly Weather Review*, submitted, 2015.

[32] M. Morzfeld, X. Tu, E. Atkins, and A.J. Chorin. A random map implementation of implicit filters. *Journal of Computational Physics*, 231:2049–2066, 2012.

[33] M. Morzfeld, X. Tu, J. Wilkening, and A.J. Chorin. Parameter estimation by implicit sampling. *Communications in Applied Mathematics and Computational Science*, submitted, 2015.

[34] T.A. Moselhy and Y.M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231:7815–7850, 2012.

[35] A. B. Owen. *Monte Carlo theory, methods and examples.* 2013.

[36] Y. Pokern, A.M. Stuart, and P. Wiberg. Parameter estimation for partially observed hypoelliptic diffusions. *J. Roy. Statis. Soc. B*, 71(1):49–73, 2009.

[37] C.E. Rasmussen and C.K.I. Williams. *Gaussian processes for machine learning.* MIT Press, 2006.

[38] N. Recca. *A new methodology for importance sampling.* Masters Thesis, Courant Institute of Mathematical Sciences, New York University, 2011.

[39] A. Samson and M. Thieullen. A contrast estimator for completely or partially observed hypoelliptic diffusion. *Stochastic Process. Appl.*, 122(7):2521–2552, 2012.

[40] C. Snyder. Particle filters, the "optimal" proposal and high-dimensional systems. *Proceedings of the ECMWF Seminar on Data Assimilation for Atmosphere and Ocean.*, 2011.

[41] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629–4640, 2008.

[42] C. Snyder, T. Bengtsson, and M. Morzfeld. Performance bounds for particle filters using the optimal proposal. *Monthly Weather Review*, submitted, 2015.

[43] A.M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[44] M.K. Tippet, J.L. Anderson, C.H. Bishop, T.M. Hamil, and J.S. Whitaker. Ensemble square root filters. *Monthly Weather Review*, 131:1485–1490, 2003.

[45] P.J. van Leeuwen. Particle filtering in geophysical systems. *Monthly Weather Review*, 137:4089–4114, 2009.

[46] P.J. van Leeuwen. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136(653):1991–1999, 2010.

[47] E. Vanden-Eijnden and J. Weare. Rare event simulation and small noise diffusions. *Communications on Pure and Applied Mathematics*, 65:1770–1803, 2012.

[48] E. Vanden-Eijnden and J. Weare. Data assimilation in the low noise, accurate observation regime with application to the Kuroshio current. *Monthly Weather Review*, 141:1822–1841, 2013.

[49] B. Weir, R. N. Miller, and Y.H. Spitz. Implicit estimation of ecological model parameters. *Bulletin of Mathematical Biology*, 75:223–257, 2013.

[50] D.S. Wilks. Effects of stochastic parameterizations in the Lorenz'96 model. *Q. J. R. Meteorol. Soc.*, 131:389–407, 2005.

[51] V.S. Zaritskii and L.I. Shimelevich. Monte Carlo technique in problems of optimal data processing. *Automation and Remote Control*, 12:95 – 103, 1975.